Detection of SNP-SNP Interaction based on the Generalized Particle Swarm Optimization Algorithm

Changyi Ma, BS[1] Junliang Shang, PhD[1]* Shengjun Li, MS[1] Yan Sun, MS[1]
School of Information Science and Engineering
Qufu Normal University
Rizhao 276826, China
xhyhforever@sina.com, shangjunliang110@163.com, qfnulsj@163.com, sunyan225@126.com

Abstract—**Most of complex diseases are believed to be mainly caused by epistatic interactions of pair single nucleotide polymorphisms (SNPs), namely, SNP-SNP interactions. Though many works have been done for the detection of SNP-SNP interactions, the algorithmic development is still ongoing due to their mathematical and computational complexities. In this study, we proposed a method, PSOMiner, based on the generalized particle swarm optimization algorithm, with mutual information as its fitness function, for the detection of SNP-SNP interaction that has the highest pathogenic effect in a SNP data set. Experiments of PSOMiner are performed on six simulation data sets under the criteria of detection power. Results demonstrate that PSOMiner is promising for the detection of SNP-SNP interaction. In addition, the application of PSOMiner on a real age-related macular degeneration (AMD) data set provides several new clues for the exploration of AMD associated SNPs that have not been described previously. PSOMiner might be an alternative to existing methods for detecting SNP-SNP interactions.**

Keywords—SNP-SNP Interaction; Mutual Information; Particle Swarm Optimization

I. Introduction

Complex diseases threatening human health, such as cancer, heart disease, and diabetes, account for about 80% of current clinical diseases. Research of complex diseases thus becomes one of the hottest fields of bioinformatics. More recently, increasing attentions of researching complex diseases have been focused on genome-wide association studies (GWAS). GWAS identify massive amounts of single nucleotide polymorphisms (SNPs) associated with complex diseases from genome-wide SNPs, which implies that GWAS are important for the investigation of complex diseases. However, these identified SNPs cannot explain the underlying mechanisms of complex diseases perfectly, one reason of which is believed to be the easily overlooked pathogenic effects of SNP-SNP interactions [1, 2]. These SNP-SNP interactions can provide valuable information related of complex diseases [3, 4]. Though many works have been done for the detection of SNP-SNP interactions, the algorithmic development is still ongoing due to their mathematical and computational complexities: the evaluation measure that determines how well a SNP-SNP interaction contributes to the phenotype; the complexity of genetic architecture of a complex disease that leads to prior knowledge unavailable, such as the order and the effect magnitude of each SNP-SNP interaction; the intensive computational burden imposed by the enormous search space, which has significant implications for GWAS.

Swarm intelligence algorithms, such as ant colony optimization algorithm, and particle swarm optimization (PSO) algorithm, might be good ways for solving these complex problems [5-9]. Among them, the PSO algorithm has several merits. First, the rules of PSO algorithm are simple. The PSO algorithm has been widely used in solving the optimal solution problems. Second, the convergence speed of the PSO algorithm is fast, and many measures in the algorithm can be used jointly to avoid falling into local optimum solution. Third, the selection of parameters has a mature theoretical research. Wu *et al.* [5] proposed a PSO based method to analyze the SNP-SNP interaction associated with hypertension. Chuang *et al.* [6] used the Gauss PSO algorithm to detect and identify the best protective association with breast cancer. These two methods are the exploration of incorporating PSO algorithm into the detection of SNP-SNP interactions. However, they only focus on finding the best genotype-genotype of a SNP-SNP interaction among possible genotypes of SNP combinations, but not the SNP-SNP interactions among possible SNP combinations. Obviously, the limited sample size of SNP data affects their computational accuracies of fitness functions and hence hinders their further applications. Furthermore, these methods are experimented on very small scale data sets (<30 SNPs) of certain complex diseases, performance of which on various kinds of large scale data sets are still unclear.

In this study, we proposed a method, PSOMiner, based on the generalized PSO algorithm with mutual information as its fitness function to detect SNP-SNP interaction that has the highest pathogenic effect in a SNP data set. PSOMiner has four stages, namely, population initialization, fitness evaluation, updating the speed and the location of each particle, Updating individual experience of each particle and the common experience of the swarm. In the stage of population initialization, SNP data are mapped as a matrix, and several parameters, including particle number, iteration number, the initial speed and location of each particle, and so on, are initialized. In the stage of fitness evaluation, the measure of mutual information is employed to compute contribution of each SNP combination to the phenotype. In the third stage, the speed and the location of each particle are updated according to its current fitness value and historical fitness value. In the final stage, individual experience of each particle and the common experi-

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

ence of the swarm are updated. Experiments of PSOMiner are performed on six simulation data sets under the criteria of detection power. Results demonstrate that PSOMiner is promising for the detection of all simulated SNP-SNP interaction models. PSOMiner is also applied on data set of age-related macular degeneration (AMD). Results show the strength of PSOMiner on real applications, and capture important features of genetic architecture of AMD, which provides new clues for biologists on the exploration of AMD associated SNPs. PSOMiner might be an alternative to existing methods for detecting SNP-SNP interactions.

II. Methods

A. Particle Swarm Optimization (PSO)

The PSO algorithm was proposed in 1995, which is based on a robust theory of swarm intelligence to search for the optimal solutions of various kinds of large scale problems [7]. The swarm intelligence describes an automatically evolving system through the simulation of the social behavior of organisms and their knowledge sharing. Valuable information can be shared in the swarm to offer a certain objective which leads individuals toward optimal results [6].

In the PSO algorithm, a particle represents a possible solution. In each generation, whether speed and location of a particle are updated or maintained depends on three parts: its current speed, its previous experience, and the common experience of the swarm. The location of each particle is estimated by a fitness function for providing a good search direction. Specifically, the individual experience of each particle is updated while fitness value of its current location is higher than that of its previous experience; the common experience of the swarm is updated by the one of individual experiences of all particles with the highest fitness value while such value is higher than that of the previous common experience. Based on these strategies, the swarm gradually converges to exploit the optimal solution in the final. This superior property makes the PSO algorithm become one of the popular swarm intelligence algorithms and has been applied in several fields.

Let $pbest_i$ representing the previous experience of the i_{th} particle, which can be defined as $pbest_i = (p_{i1}, p_{i2}, \cdots$, where D is the considered dimension of the solution space. Let $gbest$ representing the common experience of the swarm, that is, the best individual experience among those of all particles, which can be written as $gbest = (g_1, g_2, \cdots$. Let the location of the i_{th} particle being $x_i = (x_{i1}, x_{i2}, \cdots$, where $x \in \{1, \cdots$, and M is the number of SNPs in the data set. The i_{th} particle speed can be denoted as $v_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$, where the range of v is from $1-M$ to $M-1$.

B. PSOMiner: Application of the PSO algorithm on the detection of SNP-SNP interaction

PSOMiner is developed based on the generalized PSO algorithm, with mutual information as its fitness function, for the detection of SNP-SNP interaction that has the highest pathogenic effect in a SNP data set. Fig. 1 is the flowchart of PSOMiner, which shows that PSOMiner has four stages.

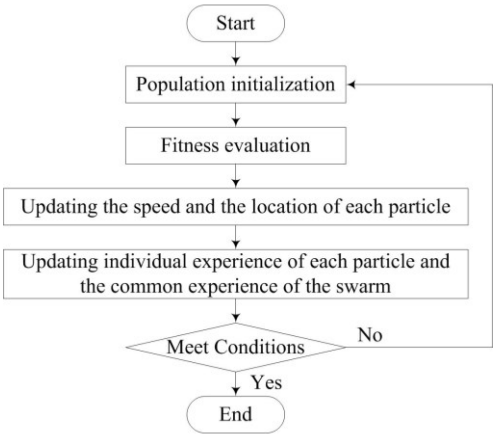
Start
Population initialization
Fitness evaluation
Updating the speed and the location of each particle
Updating individual experience of each particle and the common experience of the swarm
Meet Conditions
No
Yes
End

Fig.1. The flowchart of PSOMiner

(1) Population initialization

At present, the population way of mapping SNPs is to collect them as a matrix (Fig.2), where a row represents genotypes of a sample and a column represents a SNP. Genotypes of a SNP are coded as $\{1,2,3\}$, corresponding to homozygous common genotype (e.g., AA), heterozygous genotype (e.g., Aa), and homozygous minor genotype (e.g., aa), respectively. The label of a sample is a binary phenotype being either 1 (case) or 2 (control) [3].

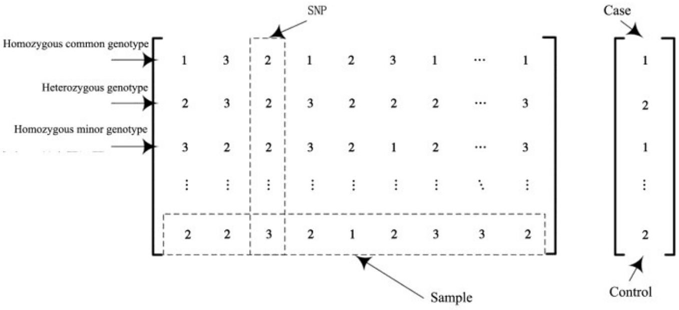

Fig.2. The SNP mapping

Here, the location of the i_{th} particle is defined as a combination of the selected SNPs, which can be defined as

$$x_i = (SNP_1, SNP_2, \cdots, SNP_m),$$

where SNP represents the index of a selected SNP, m is the considered order of the SNP-SNP interaction (in the study, it is set to 2), and $i \in \{1, \cdots$, P is the number of particles.

(2) Fitness evaluation

Fitness function of the PSOMiner plays an important role on deciding which SNP combination is the SNP-SNP interaction, and measuring how much the effect of a captured SNP-SNP interaction to the phenotype is. In the PSOMiner, mutual information is applied as its fitness function, since it is well develop and can measure multivariate dependence without

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

complex modeling. Mutual information has been widely used as a promising measure for feature selection, and is defined as

$$I(X;Y)=H(X)+H(Y)-H(X,Y),$$

where $H(X)$ is the entropy of SNP combination X; $H(Y)$ is the entropy of the phenotype Y; $H(X,Y)$ is the joint entropy of both X and Y. It is clear that higher mutual information value, namely, fitness value, indicates stronger association between the phenotype and the SNP combination.

(3) Updating the speed and the location of each particle

PSOMiner executes a search for SNP-SNP interactions by continuously updating particle speeds and particle locations in all generations. The equations for updating the speed and the location of the i_{th} particle can be defined as

$$v_{id}^{new}=w\cdot v_{id}^{old}+c_1\cdot r_1\cdot\left(pbest_{id}-x_{id}^{old}\right)+c_2\cdot r_2\cdot\left(gbest_d-x_{id}^{old}\right),$$

$$x_{id}^{new}=x_{id}^{old}+v_{id}^{new},$$

where r_1 and r_2 are random numbers between 0 and 1; learning factors c_1 and c_2 control how far a particle moving in one generation; $d\in\{1,2,\cdots$; v_{id}^{new} and v_{id}^{old} respectively denote the new and the old speeds of the i_{th} particle; x_{id}^{new} and x_{id}^{old} are respectively the new and the old locations of the i_{th} particle; the inertia weight w controls the impact of the i_{th} particle on its current speed.

(4) Updating individual experience of each particle and the common experience of the swarm

For each generation, the particle compares its fitness value of current location with that of its previous individual experience and with that of the common experience of the swarm. Both the individual experience of the particle and the common experience of the swarm are updated if the fitness value of its current location is an improvement on the previous ones. Specifically, if the fitness value of the particle location x_i is better than that of $pbest_i$, the location and fitness value of $pbest_i$ are updated to x_i. Similarly, if the fitness value of the individual experience $pbest_i$ is better than that of $gbest$, both the location and fitness value of $gbest$ are updated to $pbest_i$.

III. Results and Discussion

A. Parameter settings

Experiments of PSOMiner are performed on various simulation data sets under the criteria of detection power. To the best of our knowledge, PSOMiner is the first PSO based method for finding the SNP-SNP interaction among possible SNP combinations that has the highest pathogenic effect in a SNP data set. This is the reason of the performance of PSOMiner not being compared with that of other methods. Parameters of PSOMiner are the default settings. Specifically, particle number is set to 30; iteration number is set to 200; Inertia weight w is set to 1; learning factors c_1 and c_2 are set to 2. All experiments were performed using Matlab software in a PC environment (32-bit Windows XP system, Intel coreTM2 Quad CPU Q6600, 2.4 GHZ, 4GB RAM).

B. Detection power and data sets

Detection power is used to evaluate the performance of PSOMiner by applying PSOMiner on six groups of simulation data sets, each of which consists of a SNP-SNP interaction model. Details of these SNP-SNP interaction models are given in Table Ⅰ.

Various forms of detection power have been proposed depending on what is desired to measure [11-14]. Two types of detection power ($power1$ and $power2$) are adopted in this study with their constraints from conservative to modest [8].

Detection power 1 is defined as the proportion of data sets in which all ground-truth SNPs are detected with no false positives, where ground-truth SNPs are the SNPs in the SNP-SNP interaction models. Suppose there are N data sets with the same parameter settings and Q_i ground-truth SNPs in data set i, detection power 1 is defined as

$$power1=\sum_{i=1}^{N}\frac{x_i}{N},$$

where $x_i\in\{0,1\}$ is the detection tag, i.e., if the top Q_i SNPs detected in data set i includes all ground-truth SNPs, $x_i=1$; otherwise, $x_i=0$.

Detection power 2 is defined as an average proportion of ground-truth SNPs in the top Q_i SNPs. Detection power 2 can be written as

$$power2=\sum_{i=1}^{N}\frac{y_i}{Q_i\cdot N},$$

where y_i is the number of ground-truth SNPs in the top Q_i SNPs detected in data set i.

We exemplify six commonly used SNP-SNP interaction models (Model1 ~ Model6) for this study [11, 12, 15-17]. The first two models (Model1 and Model2) are the SNP-SNP interaction models displaying both marginal effects and interactive effects. In Model1, the penetrance increases only when both SNPs have at least one minor allele [9]. Model2 assumes that the minor allele in the first SNP has marginal effect, when minor alleles in both SNPs are present; however the effect is inversed [10]. Other models (Model3 ~ Model6) are SNP-SNP interaction models displaying no marginal effects but interactive effects. Specifically, Model3 and Model4 are directly cited from reference [11]; Model5 is a ZZ model [12] and Model6 is an XOR model [11]. Model3 ~ Model6 are exemplified since they provide a high degree of complexity to challenge ability of a method in detect pure SNP-SNP interaction effects [8]. For each model, 25 data sets are generated by simulator EpiSIM [13], where 100 SNPs are genotyped. There are 4000 individuals in each data set with 2000 cases and 2000 controls.

TABLE I. SIX SNP-SNP INTERACTION MODELS

Models	MAF (a)	MAF (b)	Preva-lence	Penetrance tables			
				A	Genotypes (B)		
					BB	Bb	bb
Model1	0.30	0.20	0.100	AA	0.087	0.087	0.087
				Aa	0.087	0.146	0.190
				aa	0.087	0.190	0.247
Model2	0.40	0.20	0.010	AA	0.009	0.009	0.009
				Aa	0.013	0.006	0.006
				aa	0.013	0.006	0.006
Model3	0.20	0.20	0.640	AA	0.486	0.960	0.538
				Aa	0.947	0.004	0.811
				aa	0.640	0.606	0.909
Model4	0.40	0.40	0.171	AA	0.068	0.299	0.017
				Aa	0.289	0.044	0.285
				aa	0.048	0.262	0.174
Model5	0.50	0.50	0.010	AA	0.000	0.020	0.000
				Aa	0.020	0.000	0.020
				aa	0.000	0.020	0.000
Model6	0.50	0.50	0.038	AA	0.000	0.000	0.100
				Aa	0.000	0.050	0.000
				aa	0.100	0.000	0.000

C. Discussion of experiment results

Detection power of PSOMiner on simulation data sets are listed on Fig.3. It shows that PSOMiner is promising for the detection of all kinds of simulated SNP-SNP interaction models. Specifically, no matter according to $power1$ or $power2$, PSOMiner always has high detection power on all models ($\geq$ 0.88); even according to $power2$, all detection power values reach to a perfect level; $power1$ on all models are higher or equal to $power2$ since $power1$ is more stricter than $power2$; detection power on Model1 are the winners among those of all models, implying that prevalence in models displaying both marginal effects and interaction effects is one reason that influences the performance of PSOMiner; detection power on those models only displaying interaction effects have relative small values, denoting that PSOMiner might sensitive to the effect type, however, this sensitivity is small, sometimes can be neglected; detection power on Model1 ~ Model2 always have different values since PSOMiner sometimes only identifies several ground-truth SNPs, but not the whole SNP-SNP interactions.

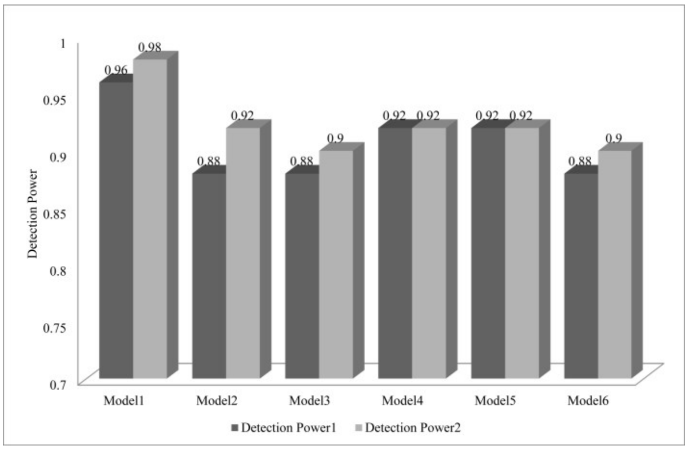

Fig.3. Detection Power of PSOMiner on simulated data sets

D. Application to real AMD data

In the study, potential of PSOMiner can also be verified by analyzing a real AMD data set [14], which contains 103611 SNPs genotyped with 96 cases and 50 controls. We run PSOMiner on AMD data set 10 times with the same parameter settings: particle number is set to 10000; iteration number is set to 1000; Inertia weight w is set to 1; learning factors c_1 and c_2 are set to 2. Detected SNP-SNP interactions associated with AMD are reported in Table Ⅱ.

The SNP-SNP interaction (rs380390, rs1374431) has the strongest interaction effect. The former one and rs1329428 are believed to be significantly associate with AMD [9], and there are biologically plausible mechanisms for the involvement of such two SNPs in AMD. The second one is located in a noncoding region. Although no evidences were reported with this gene related to AMD, it may be a plausible candidate gene associated with AMD [15]. The SNP rs2402053 is also in the intergenic region between genes *TFEC* and *TES*. It is worth noting that mutations in these genes are revealed in patients with retinal disorders. Therefore, it might be a new genetic factor contributing to the underlying mechanism of AMD.

It is interesting that the SNP-SNP interaction (rs380390, rs1363688) were successfully detected 7 times, and by other methods [3, 16, 17], though it has moderate interaction effect. Further studies with the use of large-scale case-control samples are need to confirm whether this combination have true association with AMD. We hope that some clues could be provided for the exploration of causative factors of AMD.

TABLE II. SNP-SNP INTERACTIONS ASSOCIATED WITH AMD

Index	SNP	Gene	Times	fitness value	
				Individual	Interaction
43748 63879	rs380390 rs1374431	*CFH* *N/A*	1	**0.1412** 0.0198	**0.2955**
43748 57476	rs380390 rs2402053	*CFH* *N/A*	1	**0.1412** 0.0476	0.2949
54108 31604	rs1329428 rs9328536	*CFH* *MED27*	1	0.1218 0.0563	0.2853
43748 80178	rs380390 rs1363688	*CFH* *N/A*	**7**	**0.1412** 0.0949	0.2752

IV. CONCLUSIONS

Detection and analysis of SNP-SNP interactions are believed to be important in understanding underlying mechanism of complex diseases. In this study, we proposed a method, PSOMiner, based on the generalized particle swarm optimization algorithm, with mutual information as its fitness function, for the detection of SNP-SNP interaction that has the highest pathogenic effect in a data set. To the best of our knowledge, PSOMiner is the first PSO based method for finding the SNP-SNP interaction among possible SNP combinations. PSOMiner is evaluated on six groups of simulated data sets containing SNP-SNP interaction models. Results demonstrate that PSOMiner is promising for the detection of SNP-SNP interaction. In addition, the application of PSOMiner on a real AMD data set provides several new clues for the exploration of AMD associated SNPs that have not been described previously. PSOMiner might be an alternative to existing methods for detecting SNP-SNP interactions.

It is easily seen that PSOMiner has two merits. First, PSOMiner based on a robust theory of swarm intelligence is easy to be implemented, has high capability and good generality. Second, mutual information is effective in measuring effects of SNP-SNP interactions to the phenotype. Though PSOMiner is a beneficial exploration in the detection of SNP-SNP interactions, it still has several limitations, for example, multiple SNP-SNP interactions in a data set are not considered simultaneously, PSOMiner is sensitive to those SNPs displaying strong main effects, which inspire us to continue working in the future.

Acknowledgment

This work was supported by the Science and Technology Planning Project of Qufu Normal University (xkj201410), the Scientific Research Foundation of Qufu Normal University (BSQD20130119), the Innovation and Entrepreneurship Training Project for College Students of Qufu Normal University (2014A096), the Shandong Provincial Natural Science Foundation (ZR2013FL016), the China Postdoctoral Science Foundation Funded Project (2014M560264), the Major Research Plan of the National Natural Science Foundation of China (91130006), the Specialized Research Fund for the Doctoral Program of Higher Education (20130203110017).

References

[1] L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nature Reviews Genetics,* vol. 2, pp. 91-99, 2001.

[2] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science,* vol. 273, pp. 1516-1517, 1996.

[3] J. Shang, *et al.*, "EpiMiner: A three-stage co-information based method for detecting and visualizing epistatic interactions," *Digital Signal Processing,* vol. 24, pp. 1-13, 2014.

[4] B. Maher, "The case of the missing heritability," *Nature,* vol. 456, pp. 18-21, 2008.

[5] S.-J. Wu, *et al.*, "Particle swarm optimization algorithm for analyzing SNP–SNP interaction of renin-angiotensin system genes against hypertension," *Molecular Biology Reports,* vol. 40, pp. 4227-4233, 2013.

[6] L.-Y. Chuang, *et al.*, "SNP-SNP Interaction Using Gauss Chaotic Map Particle Swarm Optimization to Detect Susceptibility to Breast Cancer," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 2014, pp. 2548-2554.

[7] K. James and E. Russell, "Particle swarm optimization," in *Proceedings of 1995 IEEE International Conference on Neural Networks*, 1995, pp. 1942-1948.

[8] J. Shang, *et al.*, "Performance analysis of novel methods for detecting epistasis," *BMC Bioinformatics,* vol. 12, p. 475, 2011.

[9] W. Tang, *et al.*, "Epistatic module detection for case-control studies: a Bayesian model with a Gibbs sampling strategy," *PLoS Genetics,* vol. 5, p. e1000464, 2009.

[10] Y. Zhang and J. S. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics,* vol. 39, pp. 1167-1173, 2007.

[11] W. Li and J. Reich, "A complete enumeration and classification of two-locus disease models," *Human Heredity,* vol. 50, pp. 334-349, 2000.

[12] W. N. Frankel and N. J. Schork, "Who's afraid of epistasis?," *Nature genetics,* vol. 14, pp. 371-373, 1996.

[13] J. Shang, *et al.*, "EpiSIM: simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis," *Genes & Genomics,* pp. 1-12, 2013.

[14] R. J. Klein, *et al.*, "Complement factor H polymorphism in age-related macular degeneration," *Science,* vol. 308, pp. 385-389, 2005.

[15] B. Han, *et al.*, "A Markov blanket-based method for detecting causal SNPs in GWAS," *BMC bioinformatics,* vol. 11, p. S5, 2010.

[16] J. Shang, *et al.*, "Incorporating heuristic information into ant colony optimization for epistasis detection," *Genes & Genomics,* vol. 34, pp. 321-327, 2012.

[17] X. Guo, *et al.*, "Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering," *BMC bioinformatics,* vol. 15, p. 102, 2014.